Genomewide demarcation of RNA polymerase II transcription units revealed by physical fractionation of chromatin

Peter L. Nagy*†, Michael L. Cleary*, Patrick O. Brown‡§, and Jason D. Lieb†‡¶‖

Departments of *Pathology and ‡Biochemistry, and §Howard Hughes Medical Institute, Stanford University, Stanford, CA 94305-5428; and ¶Department of Biology and Carolina Center for the Genome Sciences, CB #3280, 202 Fordham Hall, University of North Carolina, Chapel Hill, NC 27599-3280

Communicated by Thomas D. Petes, University of North Carolina, Chapel Hill, NC, April 3, 2003 (received for review February 10, 2003)

Epigenetic modifications of chromatin serve an important role in regulating the expression and accessibility of genomic DNA. We report here a genomewide approach for fractionating yeast chromatin into two functionally distinct parts, one containing RNA polymerase II transcribed sequences, and the other comprising noncoding sequences and genes transcribed by RNA polymerases I and III. Noncoding regions could be further fractionated into promoters and segments lacking promoters. The observed separations were apparently based on differential crosslinking efficiency of chromatin in different genomic regions. The results reveal a genomewide molecular mechanism for marking promoters and genomic regions that have a license to be transcribed by RNA polymerase II, a previously unrecognized level of genomic complexity that may exist in all eukaryotes. Our approach has broad potential use as a tool for genome annotation and for the characterization of global changes in chromatin structure that accompany different genetic, environmental, and disease states.

Genomes in eukaryotic cells contain a wealth of information not encoded directly in their DNA sequence. Three interconnected mechanisms for storing such information are well established: covalent modification of genomic DNA, most importantly methylation; alteration of chromatin by varying its protein composition; and enzymatic modification of chromatin proteins. Defects in these processes produce phenotypic effects during differentiation and development due to their profound influence on underlying gene activity (1). Histones are a major medium for such epigenetic information, because each of their tails can accommodate multiple covalent modifications, including acetylation, methylation, phosphorylation, ubiquitination, and ADP ribosylation. Specific combinations of modifications have been linked to chromatin condensation states and general transcriptional activity and may be used to guide the recruitment of transcription factors and other regulatory proteins to particular genomic regions (2–6). For example, histone H3 lysine 4 methylation by Set1p is associated with active chromatin (7), and active and repressed genes may be distinguished by di- or trimethylation of histone lysines (6). On the basis of these and other functional linkages, the information stored in histones and their modifications has been dubbed the "histone code" (8). Recently, a combination of chromatin immunoprecipitation (ChIP) and microarray techniques has allowed the genomewide distribution of histone H3 and H4 acetylated and methylated isoforms to be determined in yeast (9, 10). It is difficult, however, to assess the global effect that combinations of histone modification patterns have on the accessibility or organization of the underlying DNA template.

We report here that *Saccharomyces cerevisiae* chromatin can be fractionated physically into functionally distinct genomic regions, including coding, noncoding, and regulatory and nonregulatory domains. Two different procedures yield reciprocal results. One is based on differential segregation of untranscribed regions into the aqueous phase during phenol-chloroform extraction of formaldehyde-crosslinked chromatin. The other is a weaker enrichment for potentially transcribed regions that may occur by crosslinking-dependent genomewide nuclease protection. We propose that both fractionations are based on differential formaldehyde crosslinking of chromatin in different genomic regions, possibly mediated through differentially modified histone tail lysine residues. Our results suggest a general mechanism for demarcating regulatory and coding regions in the genome.

Materials and Methods

Data and Protocol Availability. Raw data, array images, primer sequences, compiled tabular data, detailed protocols, and additional figures are publicly available from the Stanford Microarray Database (11) (http://genome-www.stanford.edu/microarray) and the University of North Carolina Microarray Database (http://genome.unc.edu and https://genome.unc.edu/pubsup/chromatin2003).

Strains and Culture Conditions. For Experiments 1, 4, 5, and 11, strain S288C (*MAT*α *SUC2 mal mel gal2 CUP1 flo1 flo8–1*) was used. S288C was also used as a hybridization reference for experiments 3 and 28–32. For all other experiments, UCC3537 (*MAT***a** *ura3–52 lys2–801 ade2–101 leu2*-Δ1 *trp1*-Δ63 *his3*-Δ200 *adh4*::*URA3* (*URA3* at VIIL) *DIA5–1* (*ADE2* at VR), a derivative of S288C-based strain YPH250 (12), was used for both sample and reference. For all experiments, yeast was grown to an OD_{600} of 0.6−1 ($\approx 2 \times 10^8$ cells per ml) with shaking at 30°C in 50 ml of yeast extract/peptone/2% dextrose media.

Standard DNA Preparation: Experiments 1–8. All DNA was prepared by glass-bead disruption and standard phenol-chloroform extraction as described (13) with the modification that the cells were first broken in the absence of phenol-chloroform. The extract was centrifuged for 5 s at 14,000 × *g*, and the supernatant was sonicated and subsequently phenol-chloroform extracted.

Intergenic Enrichment Procedure: Experiments 9–27. Briefly, whole cells were fixed by addition of 37% formaldehyde/11% methanol (J.T. Baker) to the growth medium to a final concentration of 1% formaldehyde at 30°C for 30 min (table A at https://genome.unc.edu/pubsup/chromatin2003). Glycine was added to 125 mM from a 2.5 M stock and incubated for 5 min. The cells were centrifuged in a Sorvall RT7 at 3,000 rpm for 5 min at 4°C and washed twice with PBS and once with sterile water. Without reversing crosslinks, extracts were prepared by glass-bead disruption, sonication (fragment size 200−2,000 bp, peak at 900 bp), and standard phenol-chloroform extraction (13).

Abbreviations: ChIP, chromatin immunoprecipitation; SGD, *Saccharomyces* Genome Database; pol, polymerase.

†P.L.N. and J.D.L. contributed equally to this work.

‖To whom correspondence should be addressed. E-mail: jlieb@bio.unc.edu.

 www.pnas.org/cgi/doi/10.1073/pnas.1131966100

ORF Enrichment Procedure. Cells were crosslinked as above, and nuclei isolated as described (14) were used to prepare solubilized chromatin as described (15). Crosslinks were then reversed by incubation at 65°C, and DNA was prepared as described (5) with slight modifications. In initial experiments (nos. 28–30), immunoprecipitation using antimethyl-lysine histone H3 antibody was performed as described (5) before the crosslinks were reversed. However, the IPs were not required for ORF enrichment (Experiments 31–32).

DNA Microarrays. PCR of the individual segments and manufacture of DNA microarrays were performed as described (16). ORFs were generally represented by PCR products that extended from start codon to stop codon, regardless of intron structure. Elements representing intergenic regions included all DNA between annotated ORFs, divided such that with few exceptions, PCR products were not longer than 1.5 kb. Noncoding regions of special interest such as rDNA, tRNA, small nuclear RNA, transposon LTRs, transposons, centromeres, and introns were represented by PCR products that conformed to the *Saccharomyces* Genome Database (SGD)-annotated boundaries of specific members of each class. PCR products that represented segments of the mitochondrial genome did not necessarily conform to functional boundaries. Whole-genome primer sets can be obtained from Invitrogen (formerly Research Genetics).

Sample Amplification, Labeling, and Assay by Array Hybridization. In crosslinked samples, the DNA yield after phenol-chloroform extraction was low. Therefore, samples and references in all experiments were amplified by two initial rounds of DNA synthesis with T7 DNA polymerase (pol) by using primer A (5′-GTTTCCCAGTCACGATCNNNNNNNNN-3′), followed by 25 cycles of PCR with primer B (5′-GTTTCCCAGTCACGATC-3′) (17). Cy3-dUTP or Cy5-dUTP were then incorporated directly with an additional 25 cycles of PCR by using primer B. Microarray hybridizations were performed by using standard procedures described previously (16). Ratios were normalized by the Stanford Microarray Database default algorithm (11), and the median of pixel ratio values was retrieved for each spot. Only spots of high quality by visual inspection, with gel-verified PCR products, and whose pixels had consistent ratio values across the spot (regression correlation >0.6), were used for analysis. Arrayed elements that did not meet all of these criteria on at least half of the arrays were excluded from analysis.

Results

Physical Separation and Identification of Functionally Distinct Genomic Regions. We initially set out to investigate the global distribution of histone H3 lysine 4 methylation in yeast by ChIP. In the course of preparing genomic DNA from the formaldehyde-fixed cell extracts used in our ChIPs, we observed that if formaldehyde crosslinks were not reversed before phenol-chloroform DNA extraction, noncoding sequences were recovered in the aqueous phase with much greater efficiency than coding sequences (Fig. 1).

To confirm our initial observation, chromatin was crosslinked by addition of formaldehyde to a culture of wild-type yeast growing exponentially in rich dextrose media. Extracts from these yeast were sonicated to shear chromatin and then subjected to phenol-chloroform extraction (see *Materials and Methods*) (18). To assess the relative abundance of genomic fragments remaining in the aqueous phase, samples were RNase treated, amplified, and fluorescently labeled. In parallel, an identical procedure was used to prepare and amplify genomic DNA from noncrosslinked yeast, which was labeled with a different fluorescent marker. The two samples were then analyzed by comparative hybridization to whole-genome yeast DNA microarrays (Experiments 9–22). The microarrays contained 12,943 unique PCR products, which cover the entire yeast genome at ≈1-kb resolution (see *Materials and Methods*) (16). The microarray hybridizations revealed that during extraction of the formaldehyde-crosslinked samples, noncoding genomic regions were recovered in the aqueous phase with much greater efficiency than coding regions, consistent with our initial serendipitous observation. Comparison of the distribution of fluorescence ratio values (crosslinked/uncrosslinked) measured by ORF and intergenic spots, respectively, showed a clear separation between coding and noncoding genomic regions (Fig. 2*B*). No such separation was observed when no crosslinking was performed (Fig. 2*A*), when crosslinking was performed after the phenol-chloroform extraction step (Experiment 7), or when crosslinks were reversed before phenol-chloroform extraction (Experiment 8).

To assess the significance of the differential fractionation of individual genomic regions, for each arrayed spot we compared the ratios of fluorescence intensity in standard (Fig. 2*A*) and crosslinked (Fig. 2*B*) experiments by using a two-tailed *t* test assuming unequal variance in the two sample populations. Six thousand seven hundred thirty-three, or 52%, of the loci showed strong differential fractionation, with P values <0.05 (most of these $\ll 0.01$, Fig. 2*C*). Among these loci, 92% of the SGD-annotated ORFs had median ratios <1, whereas 87% of the elements annotated as intergenic regions had median ratios >1. Indeed, 86% (5,022 of 5,863) of all annotated ORFs for which consistent high-quality hybridization data were observed (see *Materials and Methods*) had median ratios <1. Therefore, the fate of crosslinked-chromatin fragments during phenol extraction, measured as a ratio (hybridization of crosslinked/uncrosslinked), is an excellent predictor of whether a genomic region contains an ORF (<1 ORF; >1 intergenic).

Most Putative ORFs That Fractionate Anomalously Lack Other Characteristics of Functional Genes. There is considerable evidence to suggest that many of the 262 SGD-annotated ORFs that fractionated anomalously in our assay (ratio >1 and $P < 0.05$) are not *bona fide* genes. On criteria that included size, lack of similarity to other proteins, and absence of any empirical evidence for a functional role, 43% (113/262) were independently classified as "Spurious" or "Very Hypothetical," compared with just 8% (563 of 6,368) of all ORFs classified using the same criteria (Fig. 2*D*) (19). Furthermore, SAGE tags were not isolated for 59% (156/262) of ORFs that segregated anomalously, compared with a rate of only 28% (1,709/6,208) for all SGD-annotated ORFs. Therefore, we suspect that many of the genomic fragments that were annotated as ORFs but fractionate with noncoding sequences do not actually encode proteins.

Genomic Regions Transcribed by RNA Pol I or III Segregate with Intergenic Regions. The mitochondrial genes, rDNA sequences, and sequences encoding tRNAs or small nuclear RNAs behaved very much like intergenic sequences in our assay (see Fig. 4, Experiments 9–27). Therefore, genomic regions with the potential to be transcribed by RNA pol II, but not RNA pol I, pol III, or the mitochondrial RNA pol, behaved as a distinct class in our assay.

RNA Pol II Transcription Units Are Demarcated by Chromatin Structure, Irrespective of Ongoing Transcription. Protein-coding genes that are transcriptionally silent during rapid growth in rich dextrose media (yeast extract/peptone/2% dextrose) fractionated as efficiently as those that are very heavily transcribed. For example, the dextrose-repressed genes *GAL7* and *GAL1*, in the first and fifth percentile, respectively, of genes ranked by transcription rate under these conditions, segregated similarly to RPL17A, one of the most highly transcribed genes (Fig. 1 *Lower*). By comparing fractionation efficiency to transcription rate of all

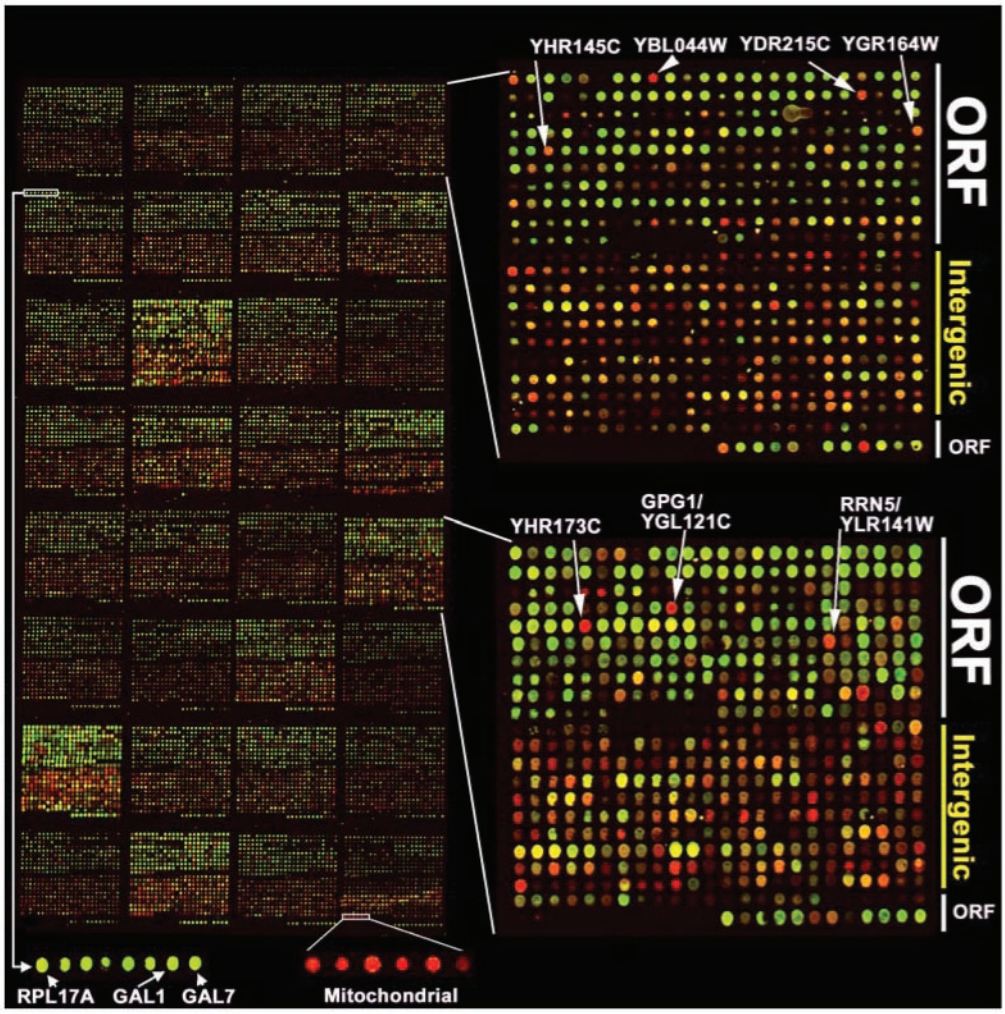

Fig. 1. Separation and detection of functionally distinct genomic regions. In general, SGD-annotated ORFs are printed on the top half of each sector, and intergenic regions are printed on the bottom half. (*Left*) DNA remaining in the aqueous fraction after phenol extraction of crosslinked extract (red) was hybridized comparatively with DNA remaining in the aqueous fraction after phenol extraction of crosslinked solubilized chromatin whose crosslinks had been reversed before extraction (green). Therefore, both intergenic enrichment (red) and ORF enrichment (green) are being assayed simultaneously (Experiment 24, jdl_g_111E). To confirm the separable enrichment observed in each channel, each sample was analyzed independently relative to a standard DNA reference (Fig. 4, Experiments 20 and 29). (*Right*) An enlarged view of two sectors. Genomic fragments corresponding to YHR145C, YBL044W, YDR250C, YGR164W, and YHR173C segregated anomalously. All are short ORFs for which there is no evidence for transcription. However, there were exceptions: arrayed elements *GPG1* and *RRN5* detected anomalous fractionation of confirmed protein-coding genes. (*Lower*) Despite differences in expression level during log-phase growth in dextrose (32), *RPL17A*, *GAL1*, and *GAL7* segregated similarly in both enrichment procedures. Mitochondrial DNA, which is nucleosome-free, was the most heavily enriched class of DNA in the crosslinked-chromatin phenol extraction procedure.

genes, we found that the efficiency of ORF fractionation did not correlate with either known RNA levels or transcription rates (Fig. 3*A*). We infer that this procedure fractionates genomic segments based on a previously undiscovered molecular characteristic of chromatin that represents a ''license'' to be transcribed by RNA pol II, rather than active transcription.

Promoter-Containing Intergenic Regions Are Distinguished from Intergenic Regions Without Regulatory Potential. The fractionation properties of intergenic regions were compared with the transcription rates of their downstream genes. Intergenic regions upstream of genes being heavily transcribed at the time of fixation were more highly enriched in the fractionation procedure than those upstream of genes with lower transcription rates (Fig. 3*B*). Moreover, intergenic regions upstream of two genes or one gene were significantly more prone to remain in the aqueous phase of the phenol-chloroform extraction than intergenic regions that do not lie directly upstream of any gene, which presumably do not contain promoters (median $\log_2$ ratios 0.34, 0.32, and 0.12, respectively, $P \ll 0.0001$ for both comparisons; see also Fig. 4).

An Alternative Physical Fractionation Procedure Results in a Reciprocal but Less Efficient Segregation of Genomic Sequences. Formaldehyde-crosslinked cells were converted to spheroplasts, permeabilized, salt extracted, and lysed to isolate nuclei. Nuclei were then disrupted by sonication to yield solubilized chromatin (see *Materials and Methods*). After reversal of crosslinks and standard phenol-chloroform preparation, microarray hybridizations revealed that genomic regions containing ORFs had been enriched in the aqueous fraction by this procedure (Fig. 4, Experiments 28–32). We discovered this effect because we attempted to use the solubilized chromatin as the starting material for unrelated ChIP experiments, but our controls revealed a selective enrichment of ORF sequences before the ChIP. The reciprocal pattern of enrichment of genomic loci in this procedure, as compared

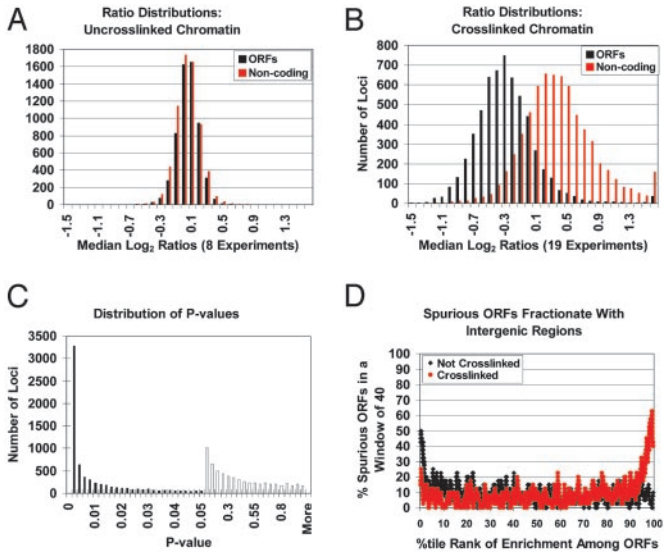

Fig. 2. ORFs and intergenic regions can be fractionated, but spurious ORFs do not segregate with functional ORFs. (*A*) Conventional phenol-chloroform extraction and DNA amplification yield unbiased DNA populations. A histogram of the distribution of normalized median $\log_2$ ratio values across eight control experiments (Experiments 1–8; see table A, https://genome.unc.edu/pubsup/chromatin2003) for SGD-annotated ORFs (black) or noncoding regions (red). All arrayed elements were plotted. (*B*) Phenol-chloroform extraction of crosslinked chromatin differentially segregates coding and noncoding regions. A histogram of the distribution of ratio medians [$\log_2$(experimental signal intensity/normalized reference signal intensity)] across 19 experiments (Experiments 9–27, table A, https://genome.unc.edu/pubsup/chromatin2003). Each DNA sample was prepared independently by phenol-chloroform extraction from crosslinked yeast, whereas reference DNA was prepared independently from noncrosslinked yeast (Experiments 9–22, for treatment in Experiments 23–27; see table A, https://genome.unc.edu/pubsup/chromatin2003). All arrayed elements were plotted: SGD-annotated ORFs (black); noncoding regions (red). (*C*) The differential segregation of individual genomic fragments using crosslinked vs. noncrosslinked chromatin. The distribution of *P* values resulting from a comparison of the ratios [$\log_2$(experimental signal intensity/normalized reference signal intensity)] at individual spots in 19 crosslinked/uncrosslinked (Experiments 9–27) and eight uncrosslinked/uncrosslinked samples (Experiments 1–8). Each bar represents the number of occurrences in increments or ''bin size'' of 0.002 for values between 0 and 0.05 is (filled bars), and in increments of 0.05 for values between 0.05 and 1 (open bars). (*D*) Annotated ORFs were ordered according to the percentile rank of their enrichment in crosslinked samples subjected to phenol/chloroform extraction, such that those to the right behave most like intergenic sequences. Plotted is the percentage of ORFs at each rank (moving window = 40, step size 1) that were classified as ''Spurious'' or ''Very Hypothetical'' (19). Many of the ORF sequences recovered from the aqueous phase are likely to be misannotated.

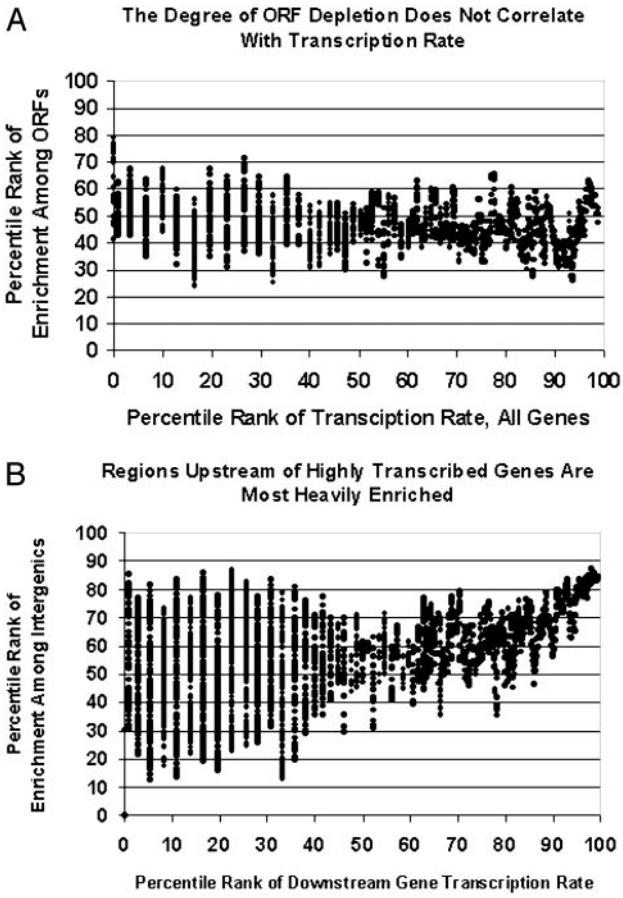

Fig. 3. The fractionation does not depend on active transcription. (*A*) The moving median (window size = 40) of percentile rank of enrichment among ORFs in 19 crosslinked-chromatin phenol extraction experiments (Experiments 9–27) is plotted against percentile rank of transcription rate (mRNAs/hr) (32). Under the tested growth conditions, there is no correlation between transcription rate and the degree of ORF depletion. See also the *GAL* genes in Fig. 1. Intron-containing genes are not included (see supplemental results at https://genome.unc.edu/pubsup/chromatin2003, for justification). (*B*) Intergenic regions upstream of heavily transcribed genes are more heavily enriched. Compare with *A*. A moving median (window = 40) of the percentile rank of enrichment among upstream intergenic regions reported in 19 crosslinked-chromatin phenol extraction experiments (Experiments 9–27) plotted against the percentile rank of the transcription rate of the downstream gene. If two genes are downstream, the highest rate is used. All upstream intergenic spots were analyzed, regardless of *P* value. The graph for data derived from intergenic regions upstream of only one gene is essentially identical (figure B at https://genome.unc.edu/pubsup/chromatin2003), showing that the observation was not created by including double promoters, which as a class are more heavily enriched than single promoters (Fig. 4) or by plotting only the most highly transcribed of two downstream genes.

with the previous procedure (Fig. 4; Experiments 9–27), suggested a mechanistic connection between the two procedures. We propose that the ORF enrichment is due to a genomewide formaldehyde-dependent nuclease protection of coding regions, as addressed in *Discussion* and Fig. 5.

Discussion

A Model for Differential Chromatin Fractionation. Local variation in chromatin composition and structure is extremely diverse and complex (8, 20), yet our studies reveal what appears to be a global pattern that systematically demarcates sequences in a way that reflects their assigned role in the transcriptional organization of the genome. The most striking and consistent feature of our results was that sequences with the potential to be transcribed by RNA pol II appear to be preferentially trapped in crosslinked chromatin (Fig. 5*A*). We suggest that when the formaldehyde-treated chromatin is fractionated by phenol extraction, crosslinked protein–DNA complexes segregate to the interphase, whereas DNA segments with fewer crosslinkable proteins, notably intergenic sequences, are preferentially released into the aqueous phase (Fig. 5*B*). Previous studies had shown that DNA yields are reduced dramatically when chromatin is crosslinked before phenol-chloroform extraction, but the composition of the DNA recovered from the aqueous phase was not analyzed (21, 22).

Conversely, when the crosslinked cell extract is incubated under conditions that allow endogenous nucleases to act, the chromatin with the highest density of crosslinked proteins is protected, whereas DNA in chromatin fragments with fewer

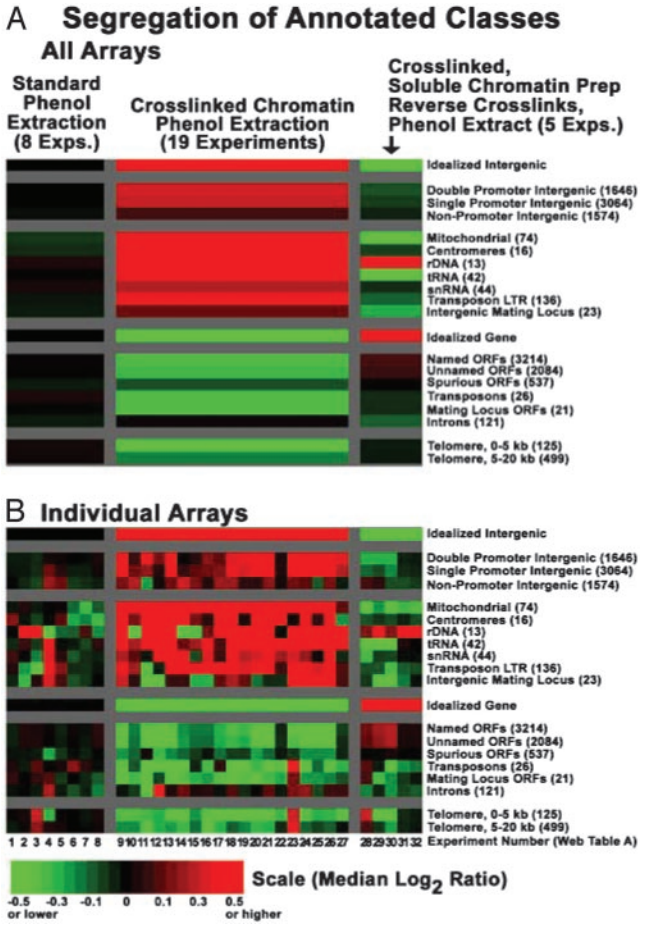

Fig. 4. Genomewide fractionation of functionally distinct genomic regions. Arrayed DNA elements were divided into functional groups based on their classification in Stanford Microarray Database, or in the case of ''Spurious ORFs'' as classified by Wood *et al.* (19) (labeled on the right). All categories are mutually exclusive, except for the telomeric classes, which contain spots that also appear in other categories. (see supplemental results at https://genome.unc.edu/pubsup/chromatin2003) The number of arrayed elements in each functional category is listed in parentheses on the right. Experiment numbers (bottom) refer to table A, https://genome.unc.edu/pubsup/chromatin2003. Experiments 1–8 compared standard genomic DNA preparations with genomic DNA preparations from (*i*) extracts that had not been crosslinked, (*ii*) extracts that had been crosslinked, followed by reversal of crosslinks, and (*iii*) extracts that had been crosslinked after phenol extraction. Experiments 9–22 revealed the intergenic enrichment phenomenon described in *Results*. In Experiments 23–27, the reference was intergenic-enriched, and the samples were ORF enriched, so the cumulative effect was measured. Experiments 28–32 revealed the ORF enrichment. See *Materials and Methods* for details of array design. (*A*) Colors (see scale) represent the median of all ratio values for all arrayed elements in each functional class (labeled on the right, top to bottom) within each of the three experimental categories (labeled at the top, left to right). (*B*) To illustrate reproducibility, medians for individual arrays, rather than across experimental categories, are shown.

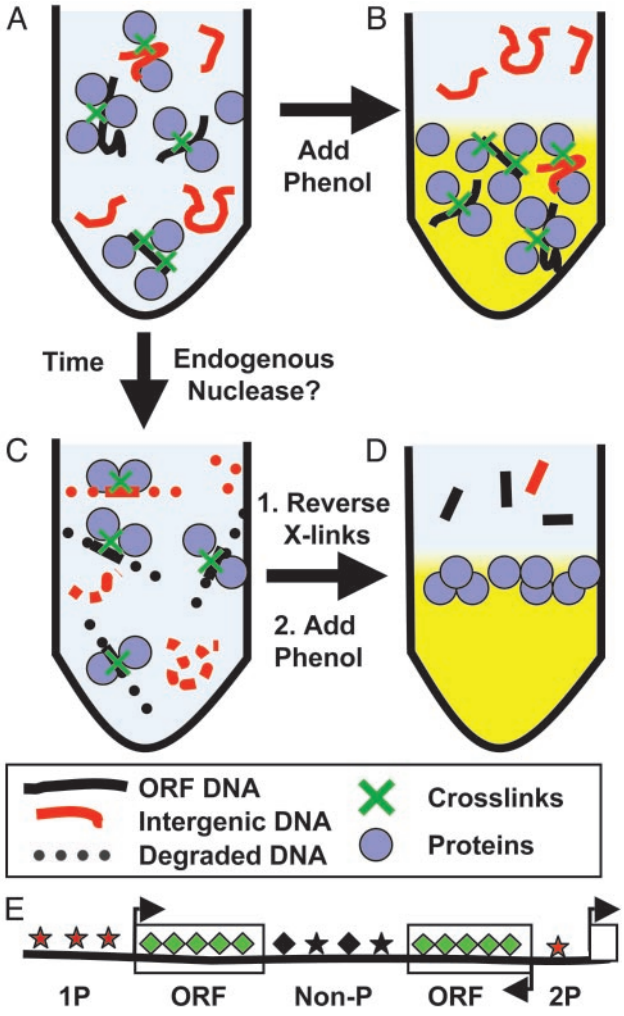

Fig. 5. Proposed mechanism for chromatin fractionation. *A–D* are described in *Discussion*. (*E*) Genomic regions upstream of two genes (2P) were most strongly enriched in the aqueous phase during the process depicted in *A* and *B*, whereas genomic regions upstream of one gene (1P) were less strongly enriched. Intergenic regions that do not contain promoters (Non-P) were neither enriched nor depleted, whereas DNA encoding ORFs was strongly depleted. The shape and spacing of the symbols associated with each genomic region represent differences in histone modification or nucleosome distribution, respectively, that may underlie the differential fractionation.

crosslinked proteins is more rapidly digested by endogenous nucleases (Fig. 5*C*). Reversal of crosslinks followed by phenol extraction reveals the differential protection of ORFs (Fig. 5*D*). Consistent with this nuclease-protection hypothesis, DNase hypersensitive sites in cellular chromatin are found predominantly in nontranscribed regions.

Formaldehyde as a Chromatin Probe. Nearly 20 years ago, Solomon and Varshavsky (23) established that the arrangement of octameric histone cores on the SV40 minichromosome is nonrandom, and that a previously identified 400-bp nuclease-hypersensitive segment of the viral genome was differentially released as naked DNA from formaldehyde-fixed, pronase, and SDS-treated minichromosome preparations. The 400-bp stretch is noncoding, contains binding sites for the T-antigen, and was interpreted to be nucleosome free. The authors suggested (23), "Thus the HCHO technique . . . could also detect and map nucleosome-free regions within cellular chromosomes *in vivo*." The most direct link between our data and the Solomon and Varshavsky study is the extremely strong segregation of nucleosome-free mitochondrial DNA to the aqueous phase in our intergenic enrichment assay. In addition, we observe that promoters upstream of heavily transcribed genes, which may be nucleosome-free or contain specific histone modifications (8, 24), enter the aqueous phase most efficiently in our intergenic enrichment assay.

Formaldehyde penetrates organic materials quickly and forms stable but reversible methylene bridges, mainly between proteins, via the ε-nitrogen atom of lysine and an adjacent amide

nitrogen of a peptide linkage (see figure C, https://genome.unc.edu/pubsup/chromatin2003) (25, 26). For DNA to react with formaldehyde, it must be partially denatured to expose the −CO-NH grouping at position 1 (N-1) of a guanine, or the exocyclic amino groups of an adenine, guanine, or cytosine (figure C, https://genome.unc.edu/pubsup/chromatin2003). Histones may be able to promote the reaction of formaldehyde with DNA due to the double-helix destabilizing effect of arginine and lysine residues (27, 28), but there is little evidence for the *in vivo* formation of direct protein–DNA crosslinks in the literature.

Over 96% of the yeast genome is thought to be nucleosomal (29), making histones by far the most abundant and readily crosslinkable protein component of chromatin (23, 30). Therefore, we suspect that our results primarily reflect heterogeneity in the distribution of nucleosomes or differential crosslinking of modified histone tail lysines to the histone octamer, neighboring octamers, or DNA (Fig. 5*E*). Two observations suggest that the fractionation is indeed based on the heterogeneity of chromatin proteins. First, the fractionation does not depend on the AT/GC content of underlying DNA, because "properly" segregating annotated ORFs have the same base composition as those that behave like intergenic regions in our assay (mean of both classes 40.5% GC, $P = 0.92$; see supplemental results at https://genome.unc.edu/pubsup/chromatin2003). Second, the fractionation properties of chromatin change as a function of distance proximal to chromosome ends (see figure D, https://genome.unc.edu/pubsup/chromatin2003), in correlation with the chromatin changes that are known to occur at telomeres (31).

This hypothesis does not imply a higher frequency of formaldehyde crosslinking in genomic regions with a higher protein density. The local formaldehyde reactivity profile of chromatin is likely to be complex, depending on the combination of reactive ε-nitrogen atoms, modifications in lysine residues, and the accessibility of those lysines to formaldehyde. Preliminary experiments indicate that methylation of histone H3 lysine 4 is not required for fractionation (P.L.N. and J.D.L., unpublished data), suggesting that multiple modifications may play a role.

Potential Applications. This method, or a similar method, may be applicable to other eukaryotic organisms, particularly if the local formaldehyde reactivity of chromatin is determined largely by conserved histone modification states. Based on its ability to differentiate true genes from noncoding regions, our method may be useful in the annotation of sequenced genomes, the creation of unbiased coding DNA libraries (analogous in use to cDNA libraries), starting material for shotgun sequencing of ORFs in highly repetitive genomes, random-clone expression microarrays for species with unsequenced genomes, and similar uses. Surgical pathologists routinely use variations in chromatin morphology observed in hematoxylin/eosin-stained tissue sections to identify specific cell types and malignancies. It is possible that a detailed genomic view of these variations may have applications in the diagnosis and subtyping of cancer and other diseases.

We thank the curators of SGD and Stanford Microarray Database for help and advice, and Dan Gottschling (Fred Hutchinson Cancer Research Center, Seattle) for strains. We thank Sandy Johnson, Joe Derisi, Vishy Iyer, and Mike Eisen for helpful discussions. The Howard Hughes Medical Institute, a fellowship from the Helen Hay Whitney Foundation (to J.D.L.), National Human Genome Research Institute Grant 1K22HG002577-01 (to J.D.L.), and National Institutes of Health Grants CA85129 (to P.O.B.), CA 34233 (to M.L.C.), and CA55029 (to M.L.C.) supported this work. P.O.B. is an investigator of the Howard Hughes Medical Institute.

1. Richards, E. J. & Elgin, S. C. (2002) *Cell* **108,** 489–500.
2. Jacobs, S. A., Taverna, S. D., Zhang, Y., Briggs, S. D., Li, J., Eissenberg, J. C., Allis, C. D. & Khorasanizadeh, S. (2001) *EMBO J.* **20,** 5232–5241.
3. Nielsen, S. J., Schneider, R., Bauer, U. M., Bannister, A. J., Morrison, A., O'Carroll, D., Firestein, R., Cleary, M., Jenuwein, T., Herrera, R. E. & Kouzarides, T. (2001) *Nature* **412,** 561–565.
4. van Leeuwen, F. & Gottschling, D. E. (2002) *Curr. Opin. Cell Biol.* **14,** 756–762.
5. Lieb, J. D., Liu, X., Botstein, D. & Brown, P. O. (2001) *Nat. Genet.* **28,** 327–334.
6. Santos-Rosa, H., Schneider, R., Bannister, A. J., Sherriff, J., Bernstein, B. E., Emre, N. C., Schreiber, S. L., Mellor, J. & Kouzarides, T. (2002) *Nature* **419,** 407–411.
7. Noma, K. I. & Grewal, S. I. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 16438–16445.
8. Strahl, B. D. & Allis, C. D. (2000) *Nature* **403,** 41–45.
9. Suka, N., Suka, Y., Carmen, A. A., Wu, J. & Grunstein, M. (2001) *Mol. Cell* **8,** 473–479.
10. Bernstein, B. E., Humphrey, E. L., Erlich, R. L., Schneider, R., Bouman, P., Liu, J. S., Kouzarides, T. & Schreiber, S. L. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 8695–8700.
11. Gollub, J., Ball, C. A., Binkley, G., Demeter, J., Finkelstein, D. B., Hebert, J. M., Hernandez-Boussard, T., Jin, H., Kaloper, M. & Matese, J. C. (2003) *Nucleic Acids Res.* **31,** 94–96.
12. Sikorski, R. S. & Hieter, P. (1989) *Genetics* **122,** 19–27.
13. Hoffman, C. S. & Winston, F. (1987) *Gene* **57,** 267–272.
14. Edmondson, D. G., Smith, M. M. & Roth, S. Y. (1996) *Genes Dev.* **10,** 1247–1259.
15. Braunstein, M., Sobel, R. E., Allis, C. D., Turner, B. M. & Broach, J. R. (1996) *Mol. Cell. Biol.* **16,** 4349–4356.
16. Iyer, V. R., Horak, C. E., Scafe, C. S., Botstein, D., Snyder, M. & Brown, P. O. (2001) *Nature* **409,** 533–538.
17. Bohlander, S. K., Espinosa, R., 3rd, Le Beau, M. M., Rowley, J. D. & Diaz, M. O. (1992) *Genomics* **13,** 1322–1324.
18. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning, A Laboratory Manual* (Cold Spring Harbor Lab. Press, Plainview, NY).
19. Wood, V., Rutherford, K. M., Ivens, A., Rajandream, M.-A. & Barrell, B. G. (2001) *Comp. Funct. Genom.* **2,** 143–154.
20. Jenuwein, T. & Allis, C. D. (2001) *Science* **293,** 1074–1080.
21. Lam, C. W., Casanova, M. & Heck, H. D. (1986) *Fund. Appl. Toxicol.* **6,** 541–550.
22. Casanova-Schmitz, M. & Heck, H. D. (1983) *Toxicol. Appl. Pharmacol.* **70,** 121–132.
23. Solomon, M. J. & Varshavsky, A. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 6470–6474.
24. Myers, F. A., Evans, D. R., Clayton, A. L., Thorne, A. W. & Crane-Robinson, C. (2001) *J. Biol. Chem.* **276,** 20197–20205.
25. Feldman, M. Y. (1973) *Prog. Nucleic Acid Res. Mol. Biol.* **13,** 1–49.
26. Gustavson, K. H. (1956) *The Chemistry of Tanning Processes* (Academic, New York).
27. Siomin, Y. A., Simonov, V. V. & Poverenny, A. M. (1973) *Biochim. Biophys. Acta* **331,** 27–32.
28. Doenecke, D. (1978) *Hoppe-Seylers Z. Physiol. Chem.* **359,** 1343–1352.
29. Polach, K. J. & Widom, J. (1995) *J. Mol. Biol.* **254,** 130–149.
30. Brutlag, D., Schlehuber, C. & Bonner, J. (1969) *Biochemistry* **8,** 3214–3218.
31. Smith, C. D., Smith, D. L., DeRisi, J. L. & Blackburn, E. H. (2003) *Mol. Biol. Cell* **14,** 556–570.
32. Holstege, F. C., Jennings, E. G., Wyrick, J. J., Lee, T. I., Hengartner, C. J., Green, M. R., Golub, T. R., Lander, E. S. & Young, R. A. (1998) *Cell* **95,** 717–728.